STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:3447–3464 (DOI: 10.1002/sim.1269)

Two approaches to mutation detection based on functional data‡

Ruth M. Pfeiffer[1,*,†], Efstathia Bura[2], Amelia Smith[1] and Joni L. Rutter[1]

[1]*National Cancer Institute, 6120 Executive Blvd, Bethesda, MD 20892-7244, U.S.A.*
[2]*Department of Statistics, George Washington University, Washington, DC 20052, U.S.A.*

SUMMARY

A new technique, denaturing high-performance liquid chromatography (dHPLC), allows for detection of any heterozygous sequence variation in a gene without prior knowledge of the precise location of the sequence change. The results of a dHPLC analysis are recorded in real-time in the form of a chromatogram that is sequence-specific. In this paper we present methods to classify an individual, based on the observed chromatogram, as a homozygous wild-type or a carrier of a specific variant for the given DNA segment by comparison to representative chromatograms that are obtained from the training set of individuals with known variant status. The first approach consists of finding a parsimonious parametric model and then classifying each newly observed curve based on comparing the most discriminating characteristic, the main mode, to the main mode of the training curves. The second approach consists of finding empirical estimates of the modes of each chromatogram and using a bootstrap test for equality with the corresponding estimates of the training curves. We apply both methods to data on the breast cancer susceptibility gene *BRCA1* and test the performance of the methods on independent samples. Published in 2002 by John Wiley & Sons, Ltd.

KEY WORDS: classification; mode; non-linear regression; superpopulation model; bootstrap

1. INTRODUCTION

Historically, DNA polymorphisms and mutations were detected by routine sequencing efforts, an approach that is costly and slow.

Denaturing high-performance liquid chromatography (dHPLC) is a relatively new, fast and inexpensive procedure for detection of any heterozygous sequence variation in a gene segment, known as an amplicon, without prior knowledge of the exact location of the sequence change. Unlike complete DNA sequencing, dHPLC does not indicate the exact base pair change, but its high sensitivity and specificity and its low cost make dHPLC one of the most powerful and increasingly popular tools for discovering and analysing genetic variation in the human

* Correspondence to: Ruth M. Pfeiffer, Division of Cancer Epidemiology and Genetics, National Cancer Institute, DCEG, 6120 Executive Blvd., EPS/8030, Bethesda, MD 20892-7244, U.S.A.
† E-mail: pfeiffer@mail.nih.gov

Received March 2001
Published in 2002 by John Wiley & Sons, Ltd. *Accepted February 2002*

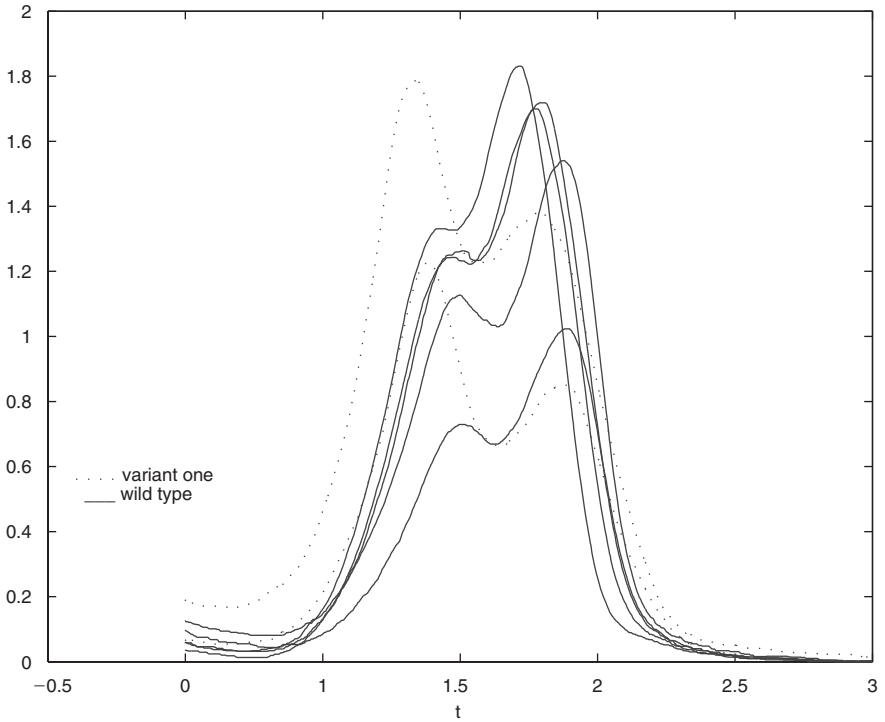

Figure 1. Training curves for amplicon 1-20.

and other genomes. Mutational analysis of candidate genes in germline and somatic mutations using dHPLC has applications to such conditions as neuroblastoma [1], prostate cancer [2], atherosclerosis [3] and ovarian tumours [4]. Other applications of dHPLC include the assembling of extensive catalogues of single nucleotide polymorphisms (SNPs) in candidate genes for particular diseases that can be used in association studies.

Figure 1 illustrates the results of a dHPLC analysis for amplicon 1-20 of the breast cancer susceptibility gene *BRCA1* for six individuals, four of whom are normal or 'wild-type', and two have a sequence change, and belong thus to the 'variant' (mutation, polymorphism) carriers. The ordinate in this figure is absorbance, which measures DNA concentration eluting from the chromatographic column, and the abscissa denotes time since the beginning of the elution process, that is, the retention time. For reasons explained in Section 2, curves corresponding to wild-type DNA have longer retention time distributions than curves based on variant DNA. Thus, based on these features of the observed chromatogram, a person can be classified as a homozygous wild-type or a carrier of a specific variant for the given DNA segment.

Currently, variants are detected 'by eye', based on a comparison with curves from a training sample of individuals with known variant status. This is a tedious, time-consuming and highly subjective process prone to variation in interpretation. In this paper we develop a robust characterization of the chromatograms and a statistical procedure that allows one to

automate the classification of the observed curves by carrier status based on comparison to the training set.

The dHPLC technique and the resulting curve data are described in more detail in Section 2. We present two approaches to classification in Sections 3 and 4, respectively. The first approach, motivated by a traditional way of analysing chromatograms [5], consists of finding a parsimonious parametric model that captures the main features of the training curves well and then classifying each newly observed curve by comparing the most discriminating characteristic, the main mode, to the main mode of the training curves. In the second approach, empirical estimates of the modes of each chromatogram are obtained, and a bootstrap test for equality with the corresponding estimates of the training curves is then carried out. In Section 5 we apply both methods to data on the breast cancer susceptibility gene *BRCA1*. We conclude in Section 6 with a discussion of our work and future directions of research.

Our work was motivated by ongoing research in the Laboratory of Population Genetics of the National Cancer Institute on genes that influence the risk of breast cancer. Mutations in *BRCA1* and *BRCA2* have been shown to predispose women to early-onset breast cancer and other malignancies. While three mutations, the so-called 'founder mutations', have been extensively studied in Ashkenazi Jewish populations, little is known about other variants in these two genes. An ongoing project seeks to identify other variants and complete the analysis of both *BRCA1/2* in a population based series of breast cancer cases from the U.S. Radiologic Technologist Health Study. This analysis will yield estimates of mutation prevalence and will permit future analyses of breast cancer risk factors in this cohort to be stratified on *BRCA1/2* status.

2. DATA DESCRIPTION

2.1. The dHPLC procedure

In dHPLC analysis a specific region of DNA, an amplicon, ranging from 200 to 700 base-pairs, is amplified by polymerase chain reaction (PCR). The product is heated to 95 degrees, to separate the DNA strands, and then slowly cooled, allowing the DNA to reanneal less stringently. Figure 2 is a schematic presentation of heteroduplex formation for a heterozygous mutation carrier. In this example, the reannealing results in the original two types of homoduplexes (A–T and G–C) and in two additional types of heteroduplexes that are mixtures of the wild-type and mutant strands (A–C and G–T). The non-Watson–Crick base pairs A–C and G–T form a 'bubble'. Each of the resulting four duplexes will pass through the dHPLC system at a different speed, and the results are recorded in real-time in the form of a chromatogram (for further detail, see, for example, reference [6]). The retention times of the heteroduplexes and homoduplexes are thus sequence-specific. They depend on the mismatched base-pairs, the nearest neighbour sequences of the mismatched base-pair [7] and the hydrogen bonding between the non-Watson–Crick paired bases [8]. In the absence of the influence of other factors, the example presented in Figure 2 would result in a curve with four distinct peaks. The peaks corresponding to the heteroduplexes would occur earlier than the peaks for the homoduplexes, because the non-Watson–Crick bases are less stable, as they are partially denatured, and thus elute from the chromatographic column faster. If an individual were homozygous, either wild-type or mutant, the reannealing would result in a single

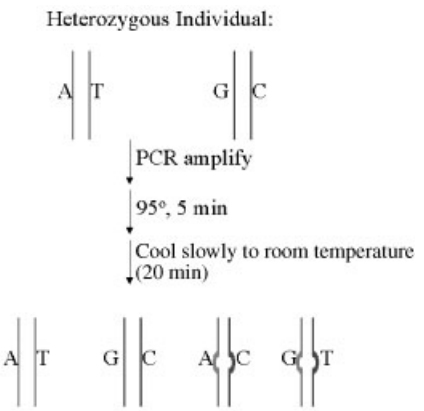

Figure 2. Heteroduplex formation.

type of homoduplex, and thus a single peak. The location of the mode of the homozygous curve is sequence-specific as well, so it is still possible to distinguish between the wild-type and the mutation carrier. In practice, though, the probability of a homozygous mutant is negligible.

In principle, one should thus be able to distinguish wild-type and variant carriers by the number of peaks observed in their chromatograms. In practice, four peaks are rarely observed, for any given DNA segment. Sequence-specific characteristics, such as the influence of the neighbouring pairs, can cause the retention times for the heteroduplexes and homoduplexes to be so close together that the result is a bimodal, or even unimodal, curve, in which case the location of the mode is shifted to the left of the wild-type modes.

There is also variation between curves of the same class, due to variations in PCR, such as concentration of DNA, melting temperature used, variation in magnesium concentration, and different primer characteristics. The absolute heights of the modes and the area under the curve are proportional to the amount of DNA used in the analysis, which may be difficult to control precisely. The relative height of the modes, compared to the highest one, and the location of modes are not much influenced by variations in DNA concentration. Our classification procedure thus focuses on these characteristics. For the *BRCA1* data, as well as for various other unreported data, the location of the maximal mode emerges as the most reliable discriminating feature.

2.2. The basic model

The data consist of chromatograms each corresponding to a subject whose PCR-amplified DNA segment was analysed using dHPLC. Following an approach discussed in Rice and Silverman [9], we consider each sample curve to arise from the model

$$Y^v(t) = \mu^v(t) + \varepsilon^v, \quad 0 \leqslant t \leqslant T \tag{1}$$

where Y denotes the absorbance, t the retention time of the chromatogram, and $E(Y^v(t))=\mu^v(t)$ is the mean of the vth variant ($v\geqslant 1$) observed in the analysed DNA segment. The instrument error is denoted by ε. The sample consists of m curves, each a realization of $Y^v(t)$ for some v, observed at times $t_1, t_2, \ldots, t_{n_i}$, $i=1,2,\ldots,m$. The observation $y_{ij}=y_i(t_j)$ denotes the jth point of the ith curve, $i=1,\ldots,m$, $j=1,\ldots,n_i$. All chromatograms are based on equidistant time points and on the same number of observed points; that is, $n_i=n$ for $i=1,2,\ldots,m$. Our goal is to classify each person in the sample as the carrier of variant v based on observation of Y^v and information available from a training data set. The total number of variants that can be observed for the analysed DNA region is not known *a priori*, in contrast to standard classification analysis, and thus the number of classes obtained from the training data is merely a lower bound. It is conceivable that a person carries a variant that was not observed in the training data. In this case, it would be appropriate to reject the hypothesis that the person belongs to any previously observed class.

A general approach to classifying the amplicons would be by comparison of their means μ^v, $v=1,2,\ldots$. This is difficult, not in the least because it is an infinite dimensional problem. Procedures for comparing mean curves rely on the global shape of the curves [10–12]. These methods are sensitive to contaminations that may influence certain regions of the curve, such as the tails, more heavily than others.

Motivated by the physical process that generates the chromatograms as described in Section 2.1, we classify the curves based on a more specific feature, the location of the modes, and in a further simplification, the location of the main mode, instead of estimating and testing for the whole mean function μ^v.

The basic classification procedure we develop is carried out as follows. First, the location and heights of the modes, for each μ^v, are estimated for curves derived from individuals with known variant status determined by sequencing. Next, the location and heights of the modes are estimated for each newly observed curve and compared to the estimates obtained from the training set. The length of the analysed DNA segments and the fact that variants are relatively rare within a segment imply that the anticipated number of different variants in each amplicon is small. In particular, in the *BRCA* genes the number of different variants does not exceed two or three for most amplicons. Nonetheless, if a new chromatogram cannot be categorized into any of the previously observed classes, sequencing the DNA is recommended to identify the true variant status and perhaps discover a new variant.

3. CLASSIFICATION OF THE CHROMATOGRAMS: A PARAMETRIC APPROACH

3.1. The individual curve model

We first estimate the location of the peaks and their associated heights based on a parametric mixture model that follows an approach presented in Robin [5]. The observed chromatogram (or thermogram, in the context of reference [5]) y is modelled as

$$y^v(t;\boldsymbol{\omega}^v)=\phi(t;\boldsymbol{\omega}^v)+\varepsilon^v \tag{2}$$

where $\boldsymbol{\omega}^v$ is a set of parameters specific to curve category v. The error term is assumed to be independent across time and normally distributed with mean zero and variance σ^2. The mean function $\phi^v=E(Y^v)$ is modelled by a mixture of several curves, each corresponding to a peak

in the chromatogram

$$\phi^v(t;\boldsymbol{\omega}^v)=\sum_{i=1}^{K^v}\varphi(t,\boldsymbol{\theta}_i^v)$$

Several choices for φ are suggested in reference [5]: the linear-Gaussian model, the Gaussian-exponential model, and the Gaussian-Gaussian model. In our problem the Gaussian-Gaussian model and, in a slight modification, the exponential-Gaussian model seem to be the best choices for a peak model based on comparison of the residual sums of squares of the models. We thus parameterize φ as

$$\varphi(t,\boldsymbol{\theta})=\begin{cases} he^{\frac{-\lambda^2(t-s)^2}{2}} & \text{if } t\leqslant s \\ he^{\frac{-\gamma^2(t-s)^2}{2}} & \text{otherwise} \end{cases}$$

for the Gaussian-Gaussian model, or

$$\varphi(t,\boldsymbol{\theta})=\begin{cases} he^{-\lambda(t-s)} & \text{if } t\leqslant s \\ he^{\frac{-\gamma^2(t-s)^2}{2}} & \text{otherwise} \end{cases}$$

for the exponential-Gaussian model with $\boldsymbol{\theta}=(h,\lambda,s,\gamma)$. The parameters have the following interpretation: h denotes the height of the peak; s the location, and λ,γ the 'thickness' of the ascending and descending parts, respectively. Notice that this model is not a mixture of probability density functions, but a non-linear regression model. For a fixed K^v we have $\boldsymbol{\omega}^v=(\boldsymbol{\theta}_1',\ldots,\boldsymbol{\theta}_{K^v}')'$. The elements of $\boldsymbol{\omega}^v$ are estimated by minimizing the sum of squares

$$\mathrm{SSE}(\boldsymbol{\omega}^v)=\sum_{j=1}^{n}(y_j-\phi(t_j,\boldsymbol{\omega}^v))^2 \tag{3}$$

Under the assumption that $\boldsymbol{\epsilon}\sim\mathrm{N}(0,\sigma^2\mathbf{I}_n)$, the resulting non-linear least squares estimator $\hat{\omega}$ of $\boldsymbol{\omega}$ is also the MLE and

$$\sqrt{n}(\hat{\boldsymbol{\omega}}-\boldsymbol{\omega})\xrightarrow{\mathscr{D}}\mathrm{N}(0,\sigma^2S^{-1}) \tag{4}$$

S is the uniform asymptotic limit of the matrix

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\partial\phi(t_j;\boldsymbol{\omega})}{\partial\boldsymbol{\omega}}\frac{\partial\phi(t_j;\boldsymbol{\omega})}{\partial\boldsymbol{\omega}'}=\frac{1}{n}\mathbf{F}'(\boldsymbol{\omega})\mathbf{F}(\boldsymbol{\omega})$$

where $\mathbf{F}(\boldsymbol{\omega})=(F_{jk}(\boldsymbol{\omega}))=(\partial\phi(t_j;\boldsymbol{\omega})/\partial\omega_k)$ is a $n\times 4K^v$ matrix of partial derivatives. For large n and under appropriate regularity conditions we have approximately

$$\hat{\boldsymbol{\omega}}^v-\boldsymbol{\omega}^v\sim\mathrm{N}(0,\sigma^2(\mathbf{F}'(\boldsymbol{\omega}^v)\mathbf{F}(\boldsymbol{\omega}^v))^{-1})$$

following reference [10]. The covariance matrix of the $4K^v$ non-linear least squares estimates $\hat{\boldsymbol{\omega}}^v$, $\mathrm{cov}(\hat{\boldsymbol{\omega}}^v)$ can then be approximated by

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\omega}}^v)=\hat{\sigma}^2(\mathbf{F}'(\hat{\boldsymbol{\omega}}^v)\mathbf{F}(\hat{\boldsymbol{\omega}}^v))^{-1} \tag{5}$$

where $(\mathbf{F}'(\boldsymbol{\omega}^v)\mathbf{F}(\boldsymbol{\omega}^v))^{-1}$ is evaluated at $\hat{\boldsymbol{\omega}}^v$, and the unbiased estimate $\hat{\sigma}^2=(n-4K^v)^{-1}\sum(y_j-\phi(t_j;\hat{\boldsymbol{\omega}}^v))^2$ is used.

To determine the number of components K^v in the non-parametric regression function, we fit a series of models with an increasing number of components and select the most plausible model by computing the Bayes information criterion (BIC). As we expect at most four peaks in each chromatogram, we restrict our models to $K^v \leqslant 4$ components, to avoid overfitting. Alternatively, one could use a likelihood ratio test statistic for model selection [5].

3.2. The superpopulation model

There are two kinds of variability in the chromatograms. The first kind is variability specific to the individual observation y_{ij}. This kind of variability represents the measurement error and is accounted for in model (2) through the error term ε. The second kind of variability is related to the whole chromatogram and can be thought of as representing curve-specific variation that arises from variations in genotyping, such as temperature, DNA concentration, or the presence of contaminants.

To account for chromatogram-specific variation, we assume that for a given variant v, the parameter $\boldsymbol{\omega}_i^v$ in model (2) for the ith curve is the realization from a superpopulation model $\boldsymbol{\omega}_i^v \sim \mathrm{N}(\boldsymbol{\omega}^v, \Sigma_0)$, with a variant-specific mean parameter $\boldsymbol{\omega}^v$ and covariance function Σ_0. Under this model the estimates $\hat{\boldsymbol{\omega}}_i^v$ have the conditional distribution $\hat{\boldsymbol{\omega}}_i^v | \boldsymbol{\omega}_i^v \sim \mathrm{N}(\boldsymbol{\omega}_i^v, \Sigma_i^v)$, where $\Sigma_i^v = \sigma^2(\mathbf{F}'(\boldsymbol{\omega}_i^v)\mathbf{F}(\boldsymbol{\omega}_i^v))^{-1}$. Unconditionally, we get that

$$\hat{\boldsymbol{\omega}}_i^v \sim \mathrm{N}(\boldsymbol{\omega}^v, \Sigma_0 + \Sigma_i^v) \tag{6}$$

To estimate the superpopulation parameters $\boldsymbol{\omega}^v$ and Σ_0 based on observations $\hat{\boldsymbol{\omega}}_i^v, \Sigma_i^v$, $i = 1, \ldots, k^v$ from the training set, several different approaches such as REML, MLE or empirical Bayes can be considered. We use the sample mean, $\hat{\boldsymbol{\omega}}^v = \sum_{i=1}^{k^v} \boldsymbol{\omega}_i^v / k^v$, and estimate $\hat{\Sigma}_0$ by

$$\hat{\Sigma}_0 = \hat{\Sigma}_\omega^{k^v} - \frac{1}{k^v} \sum_{i=1}^{k^v} \Sigma_i^v$$

where $\hat{\Sigma}_\omega^{k^v}$ denotes the sample variance of $\hat{\boldsymbol{\omega}}^v$ based on all k^v curves in the training set that fall into class v.

3.3. The parametric classification algorithm

To classify a new chromatogram, we first fit model (2) to the new curve. Let $\hat{\boldsymbol{\omega}}^v$ denote the estimated superpopulation parameters of the training curves for mutation status v and $\hat{\boldsymbol{\omega}}^{(\mathrm{new})}$ the estimated parameter vector of the new curve. The vector $\hat{\boldsymbol{\omega}}^{(\mathrm{new})}$ also has distribution (6). Since the chromatograms in the training set are independent of the newly observed curve, $\hat{\boldsymbol{\omega}}^v$ and $\hat{\boldsymbol{\omega}}^{(\mathrm{new})}$ are also independent.

In many amplicons, such as 1-11G in *BRCA1*, the wild-type curves appear unimodal while the variant, as for example variant one in Figure 3, is clearly bimodal. Figure 1 plots the training curves for amplicon 1-20 in *BRCA1*, which include five wild-type curves and two curves of variant one. Observe that for this amplicon the wild-type as well as the variant curves appear bimodal. When we plotted the estimates for the locations of both modes, we saw that they did not discriminate the curves. This can also be seen by looking at the curves directly – the location of the second largest mode of the variant one curves and the main mode of the wild-type curves are extremely close. On the other hand, when we focused

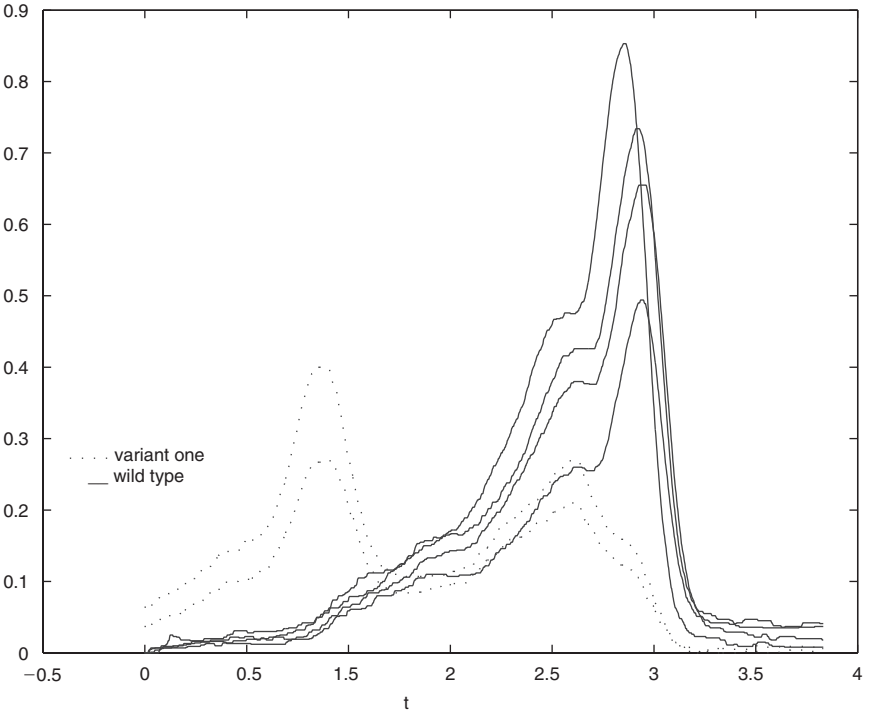

Figure 3. Training curves for amplicon 1-11G.

on the location of the main mode, the wild-type chromatograms separated well from the variant curves. We thus identify mutation status by the abscissa of the maximal mode in each curve, even when they appear to have more than one mode. The classification algorithm is easily extended to more complex features, which may be more appropriate for other amplicons.

Let $s^v_{\max}$, based on the superpopulation parameter $\boldsymbol{\omega}^v$, be the abscissa of the largest mode for mutation status $v=1,2,\ldots$, and let $s^{(\text{new})}_{\max}$ be the abscissa of the largest mode of the candidate curve whose mutation status is to be classified. For a chosen parameterization, the estimate of the abscissa of the maximal mode of the chromatogram coincides with the location component $\hat{s}_i$ of $\hat{\boldsymbol{\omega}}$ that has the largest height associated with it. Accordingly, let $\hat{s}^v_{\max}$ and $\hat{s}^{(\text{new})}_{\max}$ be their respective estimates. The hypotheses of interest are $H^v_0: s^{(\text{new})}_{\max} = s^v_{\max}$ versus $H^v_1: s^{(\text{new})}_{\max} \neq s^v_{\max}$.

Using the independence of the two estimators and (6), a test statistic for testing H^v_0 is

$$Z^v = \frac{\hat{s}^v_{\max} - \hat{s}^{(\text{new})}_{\max}}{\sqrt{\{\hat{\sigma}^2(s^v_{\max}) + \hat{\sigma}^2(s^{(\text{new})}_{\max})\}}} \tag{7}$$

If $s^v_{\max}$ denotes the lth component of $\boldsymbol{\omega}^v$, $\hat{\sigma}^2(s^v_{\max})$ is given by

$$\hat{\sigma}^2(s^v_{\max}) = \frac{1}{k^v}(\hat{\Sigma}_0)_{ll} + \frac{1}{(k^v)^2}\sum_{j=1}^{k^v}(\hat{\Sigma}_i)_{ll}$$

and $\hat{\sigma}^2(s_{\max}^{(\text{new})})$ by

$$\hat{\sigma}^2(s_{\max}^{(\text{new})}) = (\hat{\Sigma}_0)_{ll} + (\hat{\Sigma}_i^{(\text{new})})_{ll}$$

Under the null hypothesis Z^v follows a t_{k^v-1} distribution, and the rejection region is defined accordingly. As we reduce the dimensionality of the parameter space for classification, it is not necessary to estimate all elements of Σ_0, which may be problematic if there are only a few curves in the training set.

Our classification algorithm can be summarized as follows:

1. Find the best fitting model for each class in the training set and find the estimates of the location of the maximal modes based on the corresponding model. Get the superpopulation parameter estimates.
2. Fit the parametric model to the newly observed curve and get the estimates of the location of the maximal mode and its variance.
3. Starting with v = wild-type, compute the test statistic Z^v for H_0^v. If H_0^v cannot be rejected for class v in the training set, then the new observation is classified into the same category. If H_0^v is rejected, test H_0^{v+1}. A detailed description of the ordering of the hypotheses is given in the next paragraph.
4. If all H_0^vs have to be rejected, sequence the DNA to determine the true carrier status.
5. Update the training set. Classify the next curve.

In our application the training data were chosen not randomly but rather in the hope of covering as many variants as possible. For the *BRCA* genes this goal can be achieved by sequencing only breast cancer cases. If the training data were a completely random sample of the population, a Bayesian classification approach could be pursued. A prior probability of being a member of a given class could be estimated from the training data, with a subjectively chosen prior probability of being in none of the classes of the training set. The classification procedure would then consist of assigning the new observation to the class with the largest posterior probability of the largest mode, given the observed data. Even when the training set constitutes a random sample, determining the prior probability of belonging to a new class can be challenging, as for example in populations prone to admixture. The 'not seen before' class may be quite large as many different variants can be present. For the data we consider, estimation of prior probabilities is not feasible, but we still incorporate some prior knowledge about how likely membership to a given class is by ordering the H_0^vs. As the wild-type is the most common class, we start by testing whether a curve falls into this category. The other hypotheses may be tested in order of frequency of occurrence of the variants, so as to reflect some prior knowledge about the probability of class membership.

4. A BOOTSTRAP CLASSIFICATION PROCEDURE

In this section we describe a simple classification approach for the situation when only the main mode for discrimination is required. The variability specific to individual curves is fairly small, resulting in a visual impression of a roughly noiseless curve, save for the discontinuities imposed by the discrete, though very fine, time grid. Instead of fitting a parametric model to the observations, or smoothing the curves, we use the raw data to obtain an empirical estimate

of the abscissa of the main mode directly from each curve by choosing the t value that has the largest curve reading associated with it. We postulate the following model for the abscissa of the largest mode:

$$s^v_{\max,i} = s^v_{\max} + \varepsilon^v_i, \quad i = 1, \ldots, k^v$$

where v indicates the mutation status, k^v denotes the number of curves in class v in the training set and ε^v_i is a normal random variable centred at zero with variance σ_0^2. To draw the connection to the superpopulation model from the previous section, estimates $\hat{s}^v_{\max,i}$ of $s^v_{\max,i}$ have variance $\sigma_0^2 + \sigma_i^2$, where the σ_0^2 component of the variance accounts for the between-curve variation and the term σ_i^2 denotes the within-curve variation.

Let $(y^v_{i1}, y^v_{i2}, \ldots, y^v_{in})'$, $i = 1, \ldots, k^v$, represent the curves with mutation status $v = 1, 2, \ldots$, and let $\hat{s}^v_{\max,i}$ denote the observed abscissa of the largest mode of the ith curve of mutation status v. As the time grid becomes finer, that is, as $n \to \infty$, and the number of the curves of mutation status v becomes larger, that is, as $k^v \to \infty$, $\bar{s}^v_{\max} = \sum_i \hat{s}^v_{\max,i}/k^v \to s^v_{\max}$.

To test the null $H_0^v: s^{(\text{new})}_{\max} = s^v_{\max}$ at level α, we propose to find a $100(1-\alpha)$ per cent level confidence interval around the difference $d^v = s^v_{\max} - s^{(\text{new})}_{\max}$ under H_0^v and reject H_0^v if $\hat{s}^{(\text{new})}_{\max}$ falls outside this confidence interval. Unfortunately, the available number of curves per mutation status in the training set is often quite small, which makes the assumption of normality unreasonable and limits our ability to estimate the variance term. To find an approximate $100(1-\alpha)$ per cent confidence interval for d^v, we thus use the α level cut-off points of the empirical distribution function of the differences, obtained utilizing the following bootstrap approach: take a bootstrap sample of size $k^v + 1$ from the k^v curves of class v in the training set

$$y^v_{b_1 \cdot}, y^v_{b_2 \cdot}, \ldots, y^v_{b_{k^v} \cdot}, y^v_{b_{(k^v+1)} \cdot}$$

where b_i denotes the index of the ith curve in the bth bootstrap sample, and $\cdot$ stands for all points on this curve. The first k^v curves represent the bootstrap training set, and the $(k^v + 1)$th curve represents the 'new' curve under H_0. We then compute the empirical estimates for the modes; that is, we find $\hat{s}^v_{\max,b_i}$, the abscissa of the largest mode of the curve $y^v_{b_i \cdot}$, for $i = 1, \ldots, k^v + 1$. The mean for the bth bootstrap sample is based on the first k^v curves, $\bar{s}^v_{\max,b} = \sum_1^{k^v} \hat{s}^v_{\max,b_i}/k^v$, and the abscissa of the main mode of the 'new' curve is $\hat{s}^{(\text{new})}_{\max,b} = \hat{s}^v_{\max,b_{(k^v+1)}}$. Calculate $d^v_b = \bar{s}^v_{\max,b} - \hat{s}^{(\text{new})}_{\max,b}$ for $b = 1, \ldots, B$.

The bootstrap yields B estimates of the difference of the mean of the abscissae of the maximal modes in the training set and the abscissa of the maximal mode of the new curve. To test $H_0^v: s^{(\text{new})}_{\max} = s^v_{\max}$ at level α, compute the cut-off points $\hat{u}_{1-\alpha/2}$ and $\hat{u}_{\alpha/2}$ from the empirical distribution function of the bootstrapped differences $d^v_1 = \bar{s}^v_{\max,1} - \hat{s}^{(\text{new})}_{\max,1}, \ldots, d^v_B = \bar{s}^v_{\max,B} - \hat{s}^{(\text{new})}_{\max,B}$ such that

$$P(d^v_b > \hat{u}_{1-\alpha/2}) = \alpha/2$$
$$P(d^v_b \leqslant \hat{u}_{\alpha/2}) = \alpha/2$$

To classify a new observation, find the location of its largest mode $\hat{s}^{(\text{new})}_{\max}$ and check whether the observed difference between the mean of the locations of the modes of the training curves in class v and the new mode location estimate, $\hat{d}^v = \sum_i \hat{s}^v_{\max,i}/k^v - \hat{s}^{(\text{new})}_{\max}$, falls in the interval

$(\hat{u}_{\alpha/2}, \hat{u}_{1-\alpha/2})$. If it does, classify the new curve as belonging to variant status v. Otherwise, check if it falls into the confidence interval for the next class in the training set. As in the parametric approach, the hypotheses are tested in order, starting with the most likely class, the wild-type, to reflect existing prior knowledge about class membership.

If the number of curves in a class in the training set is very small, as in our examples where there are only two curves of a variant, then it is not possible to obtain exact $100(1-\alpha)$ per cent coverage for the confidence interval for this class. In this situation, under the assumption that the distributions of the differences are approximately equal for different classes, the $100(1-\alpha)$ per cent confidence interval obtained for the wild-type curves can be used to test the hypothesis for the variant class as well. This assumption is reasonable, if the variability between curves does not depend on the variant.

This bootstrap is a fast way of classifying the curves, as the confidence intervals for each class in the training set need to be calculated only once. The classification of a new observation requires only finding the abscissa of its largest mode.

5. DATA EXAMPLES

In this section we present classification results on amplicon 1-20 and amplicon 1-11G from *BRCA1* based on samples from the Radiologic Technologists Health Study. *BRCA1* consists of 35 amplicons, and *BRCA2* of 47 amplicons. All have to be analysed to get full information on the prevalence and number of different variants in these two genes. This is also true if one is interested in determining the carrier status of a single individual. An individual can be classified to be a 'variant carrier' if she has a variant in any of the 82 amplicons.

The training set for amplicon 1-20, *BRCA1*, consists of five wild-type curves and two curves corresponding to a variant that we call 'variant one'. The set does not represent a random sample from the population; it has been enriched by variant one carriers, but no inference about the prevalence of this variant in the population can be drawn. Each curve for this amplicon is based on 1501 points. The curves of the training set are plotted in Figure 1. The wild-type chromatograms as well as the variant one curves appear bimodal, and the location of the highest mode of the variant one curves is to the left of the highest mode of wild-type curves.

First, we use the parametric approach described in Section 3.3 to classify a data set that contains 49 new curves for amplicon 1-20. The DNA of every subject in this sample was sequenced as well, to determine the true carrier status. The curves in the training set are best fit with a two-component Gaussian-Gaussian model. The residual sums of squares are $0.7629, 0.8518, 0.5060, 2.6343$ and 1.2147 for the wild-type curves and 1.3769 and 0.7712 for the variant one curves. Table I shows the estimates of the maximal mode for each curve in the training set based on the Gaussian-Gaussian two-component model as well as the empirical estimate for the mode derived directly from the curve. The two estimates are very close for all curves, with the differences ranging from 0.0089 for the second wild-type curve, to -0.0580 for the last variant one curve.

As there are only two curves in the training set that represent variant one, the estimates of the superpopulation variance from these data will be poor. We thus assume that Σ_0 is the same for the wild-type and the variant one curves. This assumption is reasonable if the variability between curves does not depend on the variant.

Table I. Estimates of the location of the main mode for amplicon 1-20, *BRCA1*.

Variant	Parametric model	Empirical estimate
Wild-type	1.3042	1.2900
Wild-type	1.2223	1.2134
Wild-type	1.2817	1.2700
Wild-type	1.3994	1.3734
Wild-type	1.4053	1.3817
Variant 1	0.8204	0.8317
Variant 1	0.8889	0.8784

All the new curves are best fit with a two component Gaussian-Gaussian model as well. Table II tabulates the results for the test statistics Z^{wt} and Z^{v1} from (7) for the parametric model. The parametric approach correctly classifies all curves. One curve is classified as a variant one curve, 46 curves are wild-type curves and two curves are classified as 'neither wild-type nor variant one'. Sequencing those two curves reveals the presence of a new variant (see Table II). Figure 4 shows one of the variant two curves, together with a wild-type and a variant one curve for reference.

The values of the test statistic Z^{v} for the parametric model are quite large when the curve is not a member of variant class v. For example, $Z^{\mathrm{v1}} = -18.13$ for the first curve. This is also the case when a chromatogram is not a member of any of the classes in the training set. For example $Z^{\mathrm{wt}} = 7.97$ and $Z^{\mathrm{v1}} = -7.344$ for the last curve in the data set (Table II), which is in fact a variant two curve. In our examples, as well as in unreported data, there never was an ambiguous situation that would lead us to classify a curve as belonging to two different classes. This supports the strategy of classifying a curve as a member of variant class v if H_0^v cannot be rejected without testing all remaining hypotheses.

For the bootstrap classification, $B = 5000$ bootstrap replications were used. The 95 per cent bootstrap confidence interval for $d^v = \sum_i \hat{s}^v_{\max,i}/k^v - s^{(\mathrm{new})}_{\max}$ is $(-0.0974, 0.1010)$ for the wild-type curves. As there are only two variant one curves in the training set, the $100(1-\alpha)$ per cent confidence interval cannot be obtained, and we use the wild-type confidence interval to classify all curves. The bootstrap procedure also classifies all curves correctly. The values of $d^v = \sum_i \hat{s}^v_{\max,i}/k^v - \hat{s}^{(\mathrm{new})}_{\max}$ for each new curve are given in Table II. For the bootstrap procedure as well, the values of d^v are rather large when the new curve is not a member of class v.

For the second example, we classify curves from amplicon 1-11G, *BRCA1*. Each curve for this amplicon is based on 2000 points. Four wild-type and two variant one curves were included in the training sets. The training curves are shown in Figure 3. The variant curves are clearly bimodal, with the higher mode on the left side, while the wild-type curves have a shoulder to the left of the higher mode.

We classify 53 new curves. Again, the true carrier status for these curves is known from sequencing.

For the parametric procedure the curves in the training set are again best fit with a two-component Gaussian-Gaussian model. Table III contains the estimates of the maximal mode for each curve in the training set, based on the Gaussian-Gaussian two-component model and the corresponding empirical mode estimate. The difference between the estimates is slightly bigger than for amplicon 1-20, ranging from 0.0222 for the fourth wild-type curve to 0.0634 for the first wild-type curve.

Table II. Classification results for amplicon 1-20.

True status	Class(par)	Class(boot)	Z^{wt}	Z^{v1}	d^{wt}	d^{v1}
wt	wt	wt	0.4526	−18.1367	0.0157	−0.4427
wt*	wt	wt	0.6374	−17.8755	0.0123	−0.4461
wt	wt	wt	1.1947	−17.1118	0.0140	−0.4444
wt	wt	wt	1.2619	−17.0178	0.0173	−0.4411
wt	wt	wt	0.9270	−17.4804	0.0123	−0.4461
wt	wt	wt	0.4365	−18.1556	0.0223	−0.4361
wt	wt	wt	−0.2537	−19.1087	−0.0143	−0.4727
wt	wt	wt	1.6154	−16.5094	−0.0277	−0.4861
wt	wt	wt	−0.5685	−19.5340	−0.0643	−0.5227
wt	wt	wt	0.5533	−17.9894	−0.0510	−0.5094
wt	wt	wt	−0.0084	−18.7721	−0.0510	−0.5094
wt	wt	wt	−0.9676	−20.0922	−0.0077	−0.4661
wt	wt	wt	−1.3710	−20.6530	−0.0643	−0.5227
wt	wt	wt	−1.3628	−20.6232	−0.0393	−0.4977
wt	wt	wt	−0.3728	−19.2732	−0.0327	−0.4911
wt	wt	wt	−1.5803	−20.9355	−0.0477	−0.5061
wt	wt	wt	−0.8885	−19.9850	−0.0693	−0.5277
wt	wt	wt	−0.7486	−19.7914	−0.0293	−0.4877
wt	wt	wt	−1.1017	−20.2794	−0.0677	−0.5261
wt	wt	wt	−1.3844	−20.6499	−0.0743	−0.5327
wt	wt	wt	−0.6984	−19.7176	−0.0410	−0.4994
wt	wt	wt	−1.7905	−21.2184	−0.0193	−0.4777
wt	wt	wt	−1.9644	−21.4620	−0.0427	−0.5011
wt	wt	wt	−0.9658	−20.0881	−0.0577	−0.5161
wt	wt	wt	−0.7293	−19.7663	−0.0460	−0.5044
wt	wt	wt	−1.9068	−21.3874	−0.0710	−0.5294
v1	v1	v1	12.3074	−1.7806	0.4273	−0.0311
wt	wt	wt	−0.4972	−19.4452	−0.0427	−0.5011
wt	wt	wt	−0.9731	−20.0996	−0.0577	−0.5161
wt	wt	wt	−1.3281	−20.5732	−0.0577	−0.5161
wt	wt	wt	−0.9980	−20.1296	−0.0460	−0.5044
wt	wt	wt	−1.4942	−20.8188	−0.0710	−0.5294
wt	wt	wt	−0.4211	−19.3357	−0.0193	−0.4777
wt	wt	wt	−1.1530	−20.3479	−0.0543	−0.5127
wt	wt	wt	−0.3288	−19.2071	−0.0110	−0.4694
wt	wt	wt	−0.9557	−20.0716	−0.0343	−0.4927
wt	wt	wt	−0.3824	−19.2801	−0.0127	−0.4711
n	n	n	7.9605	−7.3443	0.2207	−0.2377
wt	wt	wt	−1.0016	−20.1370	−0.0343	−0.4927
wt	wt	wt	−1.5132	−20.8433	−0.0643	−0.5227
wt	wt	wt	−2.4054	−22.0691	−0.0877	−0.5461
wt	wt	wt	−2.4660	−22.1467	−0.0943	−0.5527
wt	wt	wt	−1.7992	−21.2372	−0.0627	−0.5211
wt	wt	wt	−1.2669	−20.5001	−0.0460	−0.5044
wt	wt	wt	−1.6169	−20.9696	−0.0660	−0.5244
wt	wt	wt	−1.5693	−20.9053	−0.0643	−0.5227
wt	wt	wt	−1.7431	−21.1315	−0.0677	−0.5261
wt	wt	wt	−1.8590	−21.2876	−0.0760	−0.5344
n	n	n	7.9605	−7.3443	0.2207	−0.2377

wt, wild-type; v1, variant 1; n, not classified; par, parametric; boot, bootstrap.

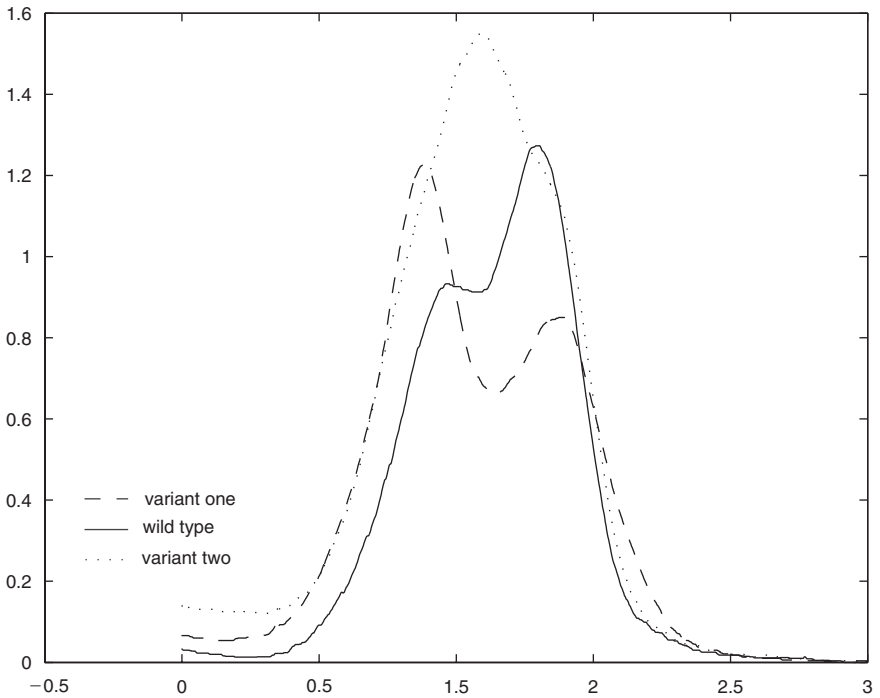

Figure 4. Curves representing all three variants observed at amplicon 1-20.

Table III. Estimates of the location of the main mode for amplicon 1-11G.

Variant	Parametric model	Empirical estimate
Wild-type	2.4768	2.4134
Wild-type	2.4610	2.4284
Wild-type	2.4695	2.4367
Wild-type	2.3739	2.3517
Variant 1	0.9369	0.8867
Variant 1	0.8963	0.8534

Similarly to amplicon 1-20, all the new curves are best fit with a two-component Gaussian-Gaussian model. In this example the parametric approach correctly classifies all but one curve. This curve, which is in fact a wild-type curve, is classified as 'neither wild-type nor variant one'. The value of the test statistic is $Z^{\mathrm{wt}} = 3.3126$, while the corresponding quantile of the t-distribution is $t_{0.975,3} = 3.1824$, and $Z^{v1} = -41.7490$ (see Table IV). Even though we reject the hypothesis that this curve is a wild-type curve, the evidence for our having truly observed a new variant is weak.

Thirteen curves are correctly identified as not belonging to any of the classes of the training set, and their sequences determine that they are members of a new class, which we call variant two. Figure 5 shows a plot of one of the chromatograms that falls into the new variant two

Table IV. Classification results for amplicon 1-11G.

True status	Class(par)	Class(boot)	Z^{wt}	Z^{v1}	d^{wt}	d^{v1}
wt	wt	wt	0.4128	−26.9215	−0.0059	−1.5434
wt	wt	wt	−0.3590	−44.8111	−0.0342	−1.5717
v2	n	n	20.7393	−27.2496	0.6175	−0.9200
wt	wt	wt	0.4531	−44.1501	0.0041	−1.5334
v2	n	n	21.2839	−26.9558	0.5975	−0.9400
wt	wt	wt	−0.5660	−44.9863	−0.0209	−1.5584
wt	wt	wt	0.5513	−44.0486	−0.0109	−1.5484
wt	wt	wt	0.4992	−44.0966	−0.0075	−1.5450
wt	wt	wt	−0.1382	−44.5939	−0.0292	−1.5667
wt	wt	wt	0.8582	−43.7848	0.0025	−1.5350
wt	wt	wt	0.9020	−43.6745	0.0091	−1.5284
wt	wt	wt	−0.0943	−44.5847	−0.0209	−1.5584
v2	n	n	23.7445	−24.9413	0.6591	−0.8784
wt	wt	wt	−0.0104	−44.5211	−0.0192	−1.5567
v2	n	n	16.3387	−28.8077	0.6108	−0.9267
wt	wt	wt	0.5160	−44.0505	−0.0175	−1.5550
v2	n	n	21.8348	−26.0054	0.5858	−0.9517
v2	n	n	22.1074	−25.9412	0.6141	−0.9234
wt	wt	wt	0.9780	−43.6335	−0.0059	−1.5434
wt	wt	wt	0.0978	−44.3976	−0.0175	−1.5550
wt	wt	wt	1.2814	−43.4810	0.0225	−1.5150
wt	wt	wt	1.4024	−43.3526	0.0325	−1.5050
wt	wt	wt	2.0188	−42.7424	0.0241	−1.5134
wt	wt	wt	2.3226	−42.4494	−0.0092	−1.5467
wt	wt	wt	0.8113	−43.8056	0.0041	−1.5334
v1	v1	v1	52.5499	−1.1439	1.4541	−0.0834
wt	wt	wt	1.7125	−43.0881	0.0391	−1.4984
wt	wt	wt	0.5257	−44.0949	−0.0025	−1.5400
wt	wt	wt	2.5726	−42.4344	0.0558	−1.4817
wt	wt	wt	1.9570	−42.8962	0.0325	−1.5050
v2	n	n	24.5316	−24.2195	0.6975	−0.8400
v2	n	n	21.5538	−25.5490	0.6858	−0.8517
v2	n	n	25.1516	−23.2852	0.7141	−0.8234
v2	n	n	24.1747	−23.7721	0.6175	−0.9200
wt	wt	wt	1.8320	−42.9102	0.0158	−1.5217
wt	wt	wt	1.9620	−42.7743	0.0308	−1.5067
wt	wt	wt	2.0222	−42.8812	0.0491	−1.4884
wt	wt	wt	2.3067	−42.6301	0.0591	−1.4784
v1	v1	v1	53.9196	−0.1591	1.5208	−0.0167
v2	n	n	23.4827	−25.1084	0.7008	−0.8367
wt	wt	wt	1.3100	−43.3950	0.0158	−1.5217
wt	wt	wt	1.0145	−43.6701	0.0041	−1.5334
v2	n	n	19.7685	−25.7300	0.7058	−0.8317
wt	wt	wt	2.1824	−42.6301	0.0225	−1.5150
v2	n	n	25.2212	−23.6839	0.7458	−0.7917
wt	wt	wt	−0.0082	−44.4704	0.0325	−1.5050
wt*	n	wt	3.3126	−41.7490	0.0625	−1.4750
wt	wt	wt	1.3818	−43.3726	0.0291	−1.5084
wt	wt	wt	1.6380	−43.1168	0.0108	−1.5267
wt	wt	wt	0.9039	−43.7673	0.0041	−1.5334
wt	wt	wt	1.1593	−43.5383	−0.0025	−1.5400
wt	wt	wt	0.5466	−44.0464	−0.0075	−1.5450
v1	v1	v1	54.2031	0.1553	1.5541	0.0166

*Marks the misclassified curve.

wt, wild-type; v1, variant 1; n, not classified; par, parametric; boot, bootstrap.

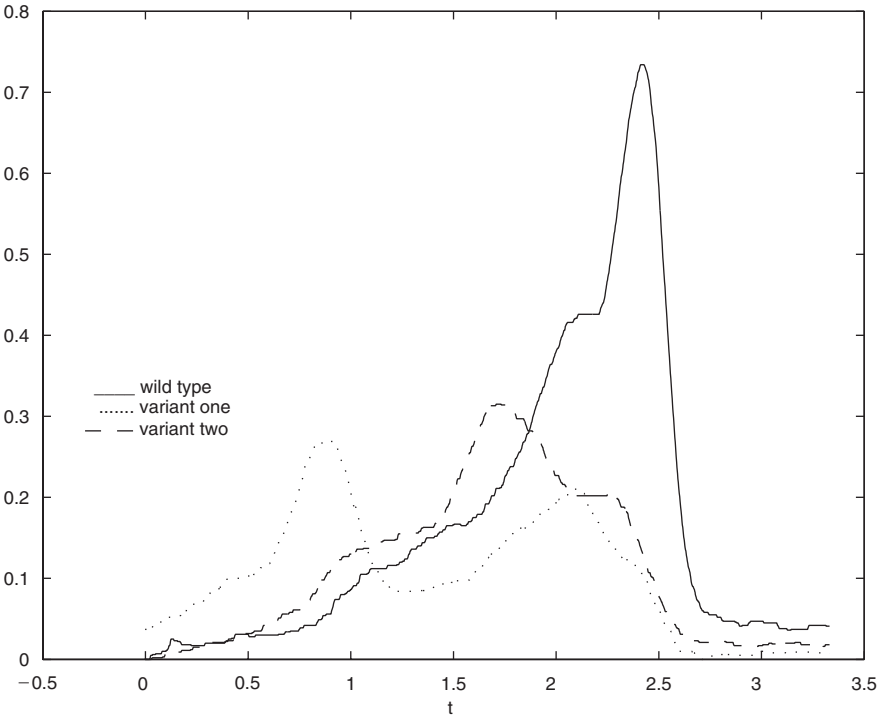

Figure 5. Curves representing all three genotypes observed at amplicon 1-11G.

class, together with a wild-type and a variant one curve for comparison. Three curves that are variant one curves, and all remaining curves that are wild-type, are correctly categorized.

For the bootstrap classification, $B=5000$ bootstrap replications were used. The 95 per cent bootstrap confidence interval for $d^v = \sum_i s^v_{\max,i}/k^v - s^{(\text{new})}_{\max}$, is $(-0.1887, 0.2097)$ for the wild-type curves. The bootstrap procedure classifies all curves correctly.

6. DISCUSSION

dHPLC is highly sensitive, with sensitivity and specificity approaching 100 per cent, has fast analysis times (less than seven minutes per run [6]), does not require post-PCR manipulation, and is favourable to pooling samples, which further lowers its already low cost per run. It has a wide range of applications, including genotyping, haplotyping and mutation detection, that are important for disease detection and analysis. In this paper we develop two methods of classifying dHPLC curves by comparing their most discriminating feature, the location of the main mode, to the same characteristic of curves with known variant status in the training set.

The first approach consists of fitting a non-linear regression model to the curves. The choice of the non-linear regression function is motivated by the fact that the main features of

chromatograms are the number and location of the peaks of the curves. Estimates of individual curve parameters are obtained via least squares. Between-curve variation is accounted for by a hierarchical model, which assumes that the individual curve parameters arise from a superpopulation. Even though we model the whole chromatogram, the classification procedure is based on comparing the location of the maximal mode of a newly observed curve to the location of the mean of the modes of each class in the training data. The location of the maximal mode was the most discriminating feature in all the chromatograms that we considered.

For the second approach we estimate the maximal modes empirically. The cut-off values for the test of equality of the modes are found through a bootstrap procedure.

By ordering the hypotheses according to the frequency of occurrence of the variants in the training set, we incorporate some prior knowledge about how likely it is that a new curve belongs to a class in the training set. As wild-type is the most likely outcome, it is the first hypothesis we test. Even if the training set were a random sample of the population, it is still somewhat difficult to assign a prior probability to the event that a curve falls into a new class. This can be especially challenging in populations that are prone to admixture, in which more variants are expected to be present.

Both methods are used to classify curves for amplicons of *BRCA1* and their performance is compared. In our examples the classification based on only the largest mode works very well, as this feature is the most discriminating one for the given amplicons, even with training sets that contain only two curves for a given variant. Only one out of 102 curves was misclassified by the parametric approach, identified as 'not in a previously observed class' when in fact it was a wild-type curve. The bootstrap testing procedure classified all curves correctly.

Amplicons for other genes may require more complex measures for discrimination. For example, if the curve appears to be bimodal with the two peaks close in height, a test based on the location of both modes may be more appropriate. The most general approach to classifying chromatograms is to test for the locations and relative heights of all observed modes. The parametric procedure easily generalizes to this situation. As the whole curve is modelled in the parametric regression approach, only the test statistic has to be adapted to a more general classification algorithm. In principle any function of the parameters can be used for discrimination of the curves. The variance of the test statistic for a function of the parameters can be found by applying the delta method. It is desirable to keep the dimensionality of the testing problem low, as the lower dimensional the testing problem is, the more powerful the test will be. It is thus recommended that one find the most parsimonious characterization that discriminates the curves well. While the parametric procedure does not require much further work to be adapted to a more complex procedure, the empirical approach becomes more involved, as different features of the curve have to be estimated from the curves. It is also harder to define a bootstrap confidence set for a multi-dimensional quantity.

Even though our examples are restricted to dHPLC curves from a single gene, *BRCA1*, chromatograms seen in many other unreported analyses are similar in shape and characteristics and our classification technique will work well for them.

The methods developed in this paper apply to chromatograms based on individual DNA samples. dHPLC may also be used to analyse pools of DNA. Future work will include developing methods to determine if a curve that is based on pooled DNA from several individuals contains one or more subjects who carry a variant and to classify the variant present in the pool.

ACKNOWLEDGEMENTS

We thank the investigators of the U.S. Radiologic Technologists Health Study at the National Cancer Institute for access to the data and Mitchell Gail, Mike Minnotte, Joseph Gastwirth, B.J. Stone and the referees for many helpful comments.

REFERENCES

1. Liu W, Mai M, Yokomizo A, Qian C, Tindall DJ, Smith DI, Thibodeau SN. Differential expression and allelotyping of the p73 gene in neuroblastoma. *International Journal of Oncology* 2000; **16**(1):181–185.
2. Seelan RS, Qian C, Yokomizo A, Bostwick DG, Smith DI, Liu W. Human acid ceramidase is overexpressed but not mutated in prostate cancer. *Genes Chromosomes and Cancer* 2000; **2**(2):137–146.
3. Cotton RGH, Bray PJ. Using CCM and DHPLC to detect mutations in the glucocorticoid receptor in atherosclerosis: a comparison. *Journal of Biochemical and Biophysical Methods* 2001; **47**(1–2):91–100.
4. Gross E, Kiechle M, Arnold N. Mutation analysis of p53 ovarian tumours by DHPLC. *Journal of Biochemical and Biophysical Methods* 2001; **47**(1–2):73–81.
5. Robin S. A model for thermograms. *Biometrics* 1999; **55**(1):37–43.
6. Kuklin A, Munson K, Gjerde D, Haefele R, Taylor P. Detection of single-nucleotide polymorphisms with the WAVE DNA fragment analysis system. *Genetic Testing* 1997; **1**(3):201–206.
7. Ke SH, Wartell RM. Influence of nearest neighbor sequence on the stability of base pair mismatches in long DNA; determination by temperature-gradient gel electrophoresis. *Nucleic Acids Research* 1993; **21**(22): 5137–5143.
8. Aboul-ela F, Koh D, Tinoco I, Martin FH. Base-base mismatches. Thermodynamics of double helix formation for dCA3XA3G+dCT3YT3G (X,Y=A.G.C.T). *Nucleic Acids Research* 1985; **13**:4811–4824.
9. Rice JA, Silverman BW. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society*, *Series B* 1991; **53**(1):233–243.
10. Hall P, Hart JD. Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association* 1990; **85**(412):1039–1049.
11. King E, Hart JD, Wehrly TE. Testing the equality of two regression curves using linear smoothers. *Statistics and Probability Letters* 1991; **12**(3):239–247.
12. Fan J, Lin SK. Tests of significance when the data are curves. *Journal of the American Statistical Association* 1998; **93**(443):1007–1021.
13. Seber GAF, Wild CJ. *Nonlinear Regression*. Wiley: New York, 1989.